To appear in: Proceedings of Lernen-Wissen-Adaption 2007 (http://lwa07.informatik.uni-halle.de/)

Evaluation of an Information Retrieval System for the Semantic Desktop using Standard Measures from Information Retrieval

Peter Scheir[1,2], **Michael Granitzer**[2], **Stefanie N. Lindstaedt**[2]
[1]Graz University of Technology, Austria
peter.scheir@tugraz.at
[2]Know-Center Graz, Austria
slind@know-center.at

Abstract

Evaluation of information retrieval systems is a critical aspect of information retrieval research. New retrieval paradigms, as retrieval in the Semantic Web, present an additional challenge for system evaluation as no off-the-shelf test corpora for evaluation exist. This paper describes the approach taken to evaluate an information retrieval system built for the Semantic Desktop and demonstrates how standard measures from information retrieval research are employed for evaluation.

1 Semantic Web information retrieval and evaluation

Despite the youthfulness of Semantic Web information retrieval, a growing amount of proposed models and implemented systems, leading from indexing triples together with textual data [Shah *et al.*, 2002], over modeling documents as parts of knowledge bases [Zhang *et al.*, 2005], to ranking search results in semantic portals [Stojanovic *et al.*, 2001], exist. Nevertheless, Semantic Web information retrieval could benefit from the experience made in information retrieval system evaluation in the past 50 years of the information retrieval discipline.

Within this paper we give an example of how standard information retrieval measures can be applied to the evaluation of retrieval performance in a Semantic Desktop environment. We aim at providing a guideline the developers of systems for the Semantic Desktop on the one hand and raise the awareness about parallels of this new domain of information retrieval to classical information retrieval on the other hand.

At present information retrieval in the Semantic Web (on the Semantic Desktop) is an inhomogeneous field (c.f. [Scheir *et al.*, 2007b]. Although a good amount of approaches does exist, different information is used for the retrieval process, different input is accepted and different output is produced. This complicates to define generally applicable rules for the evaluation of an information retrieval system for the Semantic Web (or the Semantic Desktop) and to create a test collection for this application area of information retrieval.

This paper is structured as follows: in section 2 we briefly introduce the concept of the Semantic Desktop (section 2.1) and the characteristics of our system (section 2.2). In section 3 we present the test corpus used for system evaluation. In section 4 we talk about the evaluation of the system, which measures were used (section 4.1), the queries employed for evaluation (section 4.2), how we collected relevance judgments (section 4.3) and the ranking of system configurations we have obtained (section 4.4). Finally we discuss our approach to evaluation in section 5 and conclude with section 6.

2 The evaluated system

We have built an information retrieval system for the Semantic Desktop. We will now briefly introduce the concept of the Semantic Desktop and then focus on the characteristics of the evaluated system. A detailed description of the system can be found in [Scheir *et al.*, 2007a][1]. In this paper we treat the system as a black-box and only elaborate on the input and output values of the system.

2.1 Semantic Desktop

The Semantic Desktop [Sauermann *et al.*, 2005] [Decker and Frank, 2004] paradigm stems from the Semantic Web [Berners-Lee *et al.*, 2001] movement and aims at applying technologies developed for the Semantic Web to desktop computing. In recent years the Semantic Web movement led to the development of new, standardized forms of knowledge representation and technologies for coping with them such as ontology editors, triple stores or query languages. The Semantic Desktop founds on this set of technologies and introduces them to the desktop to ultimately provide for a closer integration between (semantic) web and (semantic) desktop.

2.2 Characteristics of the evaluated system

The evaluated system relies on both, information in an ontology and the statistical information in a collection of documents. The system is queried by a set of concepts from the ontology and returns a set of documents. Documents in the system are (partly) annotated with ontological concepts if a document *deals with* a concept. For example, if the document is an introduction to use case models it is annotated with the corresponding concept in the ontology. The annotation process is performed manually but is supported by statistical techniques (e.g. identification of frequent words in the document collection) [Pammer *et al.*, 2007].

Concepts from the ontology are used as metadata for documents in the system. Opposed to classical metadata, the ontology specifies relations between the concepts. For example, class-subclass relationships are defined as well

[1]Available online under: `http://www.know-center.tugraz.at/media/files/wissensbilanz/publications_wm/papers/2007_scheir_improving_search_on_the_semantic_desktop_pdf` (01.09.2007)

as arbitrary semantic relations between concepts are modeled (e. g. `UseCase isComposedOf Action`). The structure of the ontology can be utilized for calculating the similarity between two concepts in the ontology. This similarity can be used to extend a query by similar concepts before retrieving documents dealing with a set of concepts. After retrieval of documents was performed, the result set can be extended by means of textual similarity. Different combinations of query and result expansion were evaluated against each other.

3 The test corpus

A major obstacle in the easy evaluation of Semantic Web technology based information retrieval systems is the absence of standardized test corpora, as they exist for text-based information retrieval.

Therefore we have built our own test corpus based on the data available in the first release of the APOSDLE system [Lindstaedt and Mayer, 2006]. The first version of APOSDLE was built for the domain of Requirements Engineering. This resulted into a domain ontology for this field and a set of documents dealing with various topics of Requirements Engineering. The document base was provided by a partner in the APOSDLE project, with expertise in the field of Requirement Engineering, while the ontology was modeled by another partner. Together these two partners sign responsible for the annotation of the document base with concepts from the ontology. The ontology contains 70 concepts and the document set consists of 1016 documents. 496 documents were annotated using one or more concepts. 21 concepts from the domain ontology were used to annotate documents.

In its size our test collection is comparable to test collections from early information retrieval experiments as the Cranfield or the CACM collections.

In addition to the absence of corpora for Semantic Web information retrieval we are unaware of any standard text-retrieval corpora for evaluating a system with characteristics similar to ours. We considered treating the ontological concepts used for querying our system equivalent to query terms of a text-retrieval system to be able to use a standard corpus. Therefore we would have needed some structure relating the terms contained in the documents, as it is the case with the ontology in our system which relates concepts. For this task we could have used a standard thesaurus. As this knowledge structure is different to the ontology originally used (and therefore different similarity measures had to be applied to it), this would have led us to evaluating a system with different properties than our original one.

We also considered the INEX[2] test collection for evaluating our system. INEX provides a document collection of XML documents which would have provided us with textual data associated with XML structure information. Unfortunately again an ontology relating the metadata used as XML markup is unavailable. This would have prevented us from employing (and evaluating) the functionality provided by the query expansion technique, which founds on the ontology.

[2] `http://inex.is.informatik.uni-duisburg.de/`

4 Evaluation

In this section we describe the evaluation that we performed. We talk about the evaluation measures, the queries used for evaluation, how we collected relevance judgments and about the system configuration rankings obtained.

4.1 Measures used for evaluation

The central problem in using classic IR measures as *recall* or *mean average precision* is that they require complete relevance judgments, which means that every document is judged against every query [Buckley and Voorhees, 2004]. [Fuhr, 2006] notices that recall can not be determined precisely with reasonable effort. Finally [Carterette *et al.*, 2006] states that: *Building sets large enough for evaluation of realworld implementations is at best inefficient, at worst infeasible.*

Therefore we opted for using evaluation measures that do not require hat every document is judged against every query. We decided for using precision (P) at rank 10, 20 and 30. In addition we made use of infAP [Yilmaz and Aslam, 2006] which approximates the value of average precision (AP) using random sampling.

For calculating the evaluation scores we have used the `trec_eval`[3] package, which origins from the Text REtrieval Conference (TREC) and allows for calculating a large number of standard measures for information retrieval system evaluation.

4.2 Queries used for evaluation

The queries that were used for the evaluation of the system are formed by sets of concepts.

The first version of the APOSDLE system presents resources to knowledge workers to allow them to acquire a certain competency. To realize search for resources that are appropriate to build up a certain competency, competencies are represented by sets of concepts from the domain ontology. These sets are used as queries for the search for resources. For the evaluation of the APOSDLE system all distinct sets of concepts representing competencies[4] were used as queries. In addition all concepts from the domain model not already present in the set of queries were used for evaluation purposes.

4.3 Collecting relevance judgments

8 different system configuration were tested and compared against each other based on the chosen evaluation measures. 79 distinct queries were used to query every system configuration. Queries were formed by sets of concepts stemming from the domain ontology.

For every query and system configuration the first 30 results were stored in a database table, with one row for every query-document pair. Query-document pairs returned by more than one system configuration were stored only once. The query-document pairs stored in the database-table were then judged manually by a human assessor. All query-document pairs were judged by the same person. The assessor was not involved in defining the competency to concept mappings uses as queries (c.f. section 4.2.

After relevance judgment, both, the results obtained by the different system configurations and the global relevance judgments have been stored into text files in a format appropriate for the `trec_eval` program. We then calculated the

[3] `http://trec.nist.gov/trec_eval/`

[4] different competencies can be represented by the same concepts

P(10), P(20), (P30) and infAP scores for the different system configurations.

4.4 The obtained system configuration ranking

Table 1 shows the calculated P(10), P(20), (P30) and infAP scores for the different system configurations. Table 2 shows the system configuration ranking based on the obtained evaluation scores.

Configuration 1 (conf_1) is the baseline configuration of our system. It equals a database query with a ranked list of results. Results to the query are ranked using an idf-like measure between concepts and documents.

All other configurations make use of query expansion based on semantic similarity or result expansion based on text-based similarly. Configurations 3, 4, 5, 6, 7 and 8 perform query expansion. Configurations 2, 6, 7 and 8 perform result expansion.

Configuration 4 and 5 are essentially the same approach with minor internal differences. The same holds for configuration 7 and 8.

Conf.	P(10)	P(20)	P(30)	infAP
conf_1	0.2418	0.2051	0.1700	0.1484
conf_2	0.3089	0.2778	0.2502	0.2487
conf_3	0.3165	0.2608	0.2131	0.2114
conf_4	0.3114	0.2582	0.2097	0.2001
conf_5	0.3114	0.2582	0.2097	0.2000
conf_6	0.3848	0.3405	0.3046	0.3253
conf_7	0.3924	0.3494	0.3089	0.3326
conf_8	0.3911	0.3487	0.3080	0.3318

Table 1: Evaluation scores of system configurations calculated using P(10), P(20), P(30) and infAP

Rank	P(10)	P(20)	P(30)	infAP
1 (best)	conf_7	conf_7	conf_7	conf_7
2	conf_8	conf_8	conf_8	conf_8
3	conf_6	conf_6	conf_6	conf_6
4	conf_3	conf_2	conf_2	conf_2
5	conf_4	conf_3	conf_3	conf_3
6	conf_5	conf_4	conf_4	conf_4
7	conf_2	conf_5	conf_5	conf_5
8 (worst)	conf_1	conf_1	conf_1	conf_1

Table 2: Ranking of system configurations based on P(10), P(20), P(30) and infAP

5 Discussion

We now discuss the evaluation measures used and why we think that the amount of relevance judgments collected is sufficient for a proper evaluation of our system.

5.1 P(10), P(20) and P(30)

[Buckley and Voorhees, 2000] evaluate the stability of evaluation measures. They calculate the error rate of measures based on the number of errors occurring whilst comparing two systems using a certain measure. They divide the number of errors by the total number of possible comparisons between two different systems. Based on previous research they state that an error rate of 2.9% is minimally acceptable. They find that P(30) exactly reaches this error rate of 2.9% in their experiment with 50 queries used. Finally they suggest that the amount of queries should be increased for P(n) measures, where $n < 30$. And suggest that 100 queries would be safe if the measure P(20) is used.

We performed our experiment with 79 distinct queries and used the measures P(10), P(20) and P(30). Following the results of [Buckley and Voorhees, 2000] the size of our query set should be appropriate for P(30). We are fortified in this assumption as the ranking of the 8 system configurations is identical for P(20), P(30) and infAP.

5.2 infAP

The Trec 8 Ad-Hoc collection consists of 528,155 documents and 50 queries which make a total amount of 26,407,750 possible relevance judgments. 86830 query-document relevance pairs are actually judged. This set of pairs is created by depth-100 pooling of 129 runs. Therefore 0.33% of the possible relevance judgments are performed.

Our collection consists of 1026 documents and 79 queries, which results in a total of 81,054 possible relevance judgments. This set of pairs is created by depth-30 pooling of 8 runs and 498 additional relevance judgments that were performed for runs that were not part of the experiment. 1938 query document pairs were actually judged. Therefore 2,39% of all possible relevance judgments were performed.

The depth-100 pool for the 8 evaluated runs would consist of 4138 query-document pairs. As we judged 1938 query-document pairs, we judged 46,83% of our potential depth-100 pool. [Yilmaz and Aslam, 2006] report a Kendall's tau based rank correlation of above 0.9 between infAP and AP with as little as 25% of the maximum possible relevance judgments of the depth-100 pool of the Trec 8 Ad-Hoc collection. They consider two rankings with a rank correlation of above 0.9 as equivalent.

With 46,83% of our potential depth-100 pool judged, we are confident that the infAP measure produces an estimation sufficiently accurate. Again our confidence in the results of infAP is assured by the equivalence of the ranking of the 8 system configurations for P(20), P(30) and infAP.

6 Conclusion

We have evaluated an information retrieval system for the Semantic Desktop using standard measures for information retrieval system evaluation. As classic measures for evaluation as recall and average precision require that every document is judged for every query we have chosen precision at ranks 10, 20 and 30 as evaluation measures. In addition we made use of the random sampling approach performed by the infAP measure. We are confident that our chosen approach reflects the actual relation between the system configurations as the ranking of the system configurations remains identical for the measures P(20), P(30) and infAP.

Acknowledgments

We thank the anonymous reviewers of our submission for their constructive feedback.

This work has been partially funded under grant 027023 in the IST work programme of the European Community.

The Know-Center is funded by the Austrian Competence Center program Kplus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.ffg.at/index.php?cid=95) and by the State of Styria.

References

[Berners-Lee *et al.*, 2001] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, May 2001.

[Buckley and Voorhees, 2000] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40, New York, NY, USA, 2000. ACM Press.

[Buckley and Voorhees, 2004] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, New York, NY, USA, 2004. ACM Press.

[Carterette *et al.*, 2006] Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 268–275, New York, NY, USA, 2006. ACM Press.

[Decker and Frank, 2004] Stefan Decker and Martin R. Frank. The networked semantic desktop. In *WWW Workshop on Application Design, Development and Implementation Issues in the Semantic Web*, 2004.

[Fuhr, 2006] Norbert Fuhr. Information Retrieval: Skriptum zur Vorlesung im SS 06, 19. Dezember 2006, 2006.

[Lindstaedt and Mayer, 2006] Stefanie N. Lindstaedt and Harald Mayer. A storyboard of the aposdle vision. In *Innovative Approaches for Learning and Knowledge Sharing, First European Conference on Technology Enhanced Learning, EC-TEL 2006, Crete, Greece, October 1-4, 2006*, pages 628–633, 2006.

[Pammer *et al.*, 2007] Viktoria Pammer, Peter Scheir, and Stefanie Lindstaedt. Two protégé plug-ins for supporting document-based ontology engineering and ontological annotation at document level. In *10th International Protégé Conference - July 15-18, 2007 - Budapest, Hungary*, 2007.

[Sauermann *et al.*, 2005] Leo Sauermann, Ansgar Bernardi, and Andreas Dengel. Overview and outlook on the semantic desktop. In *Proceedings of the 1st Workshop on The Semantic Desktop at the ISWC 2005 Conference*, 2005.

[Scheir *et al.*, 2007a] Peter Scheir, Chiara Ghidini, and Stefanie N. Lindstaedt. Improving search on the semantic desktop using associative retrieval techniques. In *Proceedings of I-SEMANTICS 2007 (accepted for publication)*, 2007. Available online under: `http://www.know-center.tugraz.at/media/files/wissensbilanz/publications_wm/papers/2007_scheir_improving_search_on_the_semantic_desktop_pdf` (01.09.2007).

[Scheir *et al.*, 2007b] Peter Scheir, Viktoria Pammer, and Stefanie N. Lindstaedt. Information retrieval on the semantic web - does it exist? In *LWA 2007, Lernen - Wissensentdeckung - Adaptivität, 24.-26.9. 2007 in Halle/Saale (in this volume)*, 2007.

[Shah *et al.*, 2002] Urvi Shah, Tim Finin, and Anupam Joshi. Information retrieval on the semantic web. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 461–468, New York, NY, USA, 2002. ACM Press.

[Stojanovic *et al.*, 2001] Nenad Stojanovic, Alexander Maedche, Steffen Staab, Rudi Studer, and York Sure. Seal: a framework for developing semantic portals. In *Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001), October 21-23, 2001, Victoria, BC, Canada*, pages 155–162, 2001.

[Yilmaz and Aslam, 2006] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111, New York, NY, USA, 2006. ACM Press.

[Zhang *et al.*, 2005] Lei Zhang, Yong Yu, Jian Zhou, ChenXi Lin, and Yin Yang. An enhanced model for searching in semantic portals. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 453–462, New York, NY, USA, 2005. ACM Press.